D1124–D1129 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 29 November 2012*
doi:10.1093/nar/gks1047

NPACT: Naturally Occurring Plant-based Anti-cancer Compound-Activity-Target database

Manu Mangal[1], Parul Sagar[1], Harinder Singh[2], Gajendra P. S. Raghava[2] and Subhash M. Agarwal[1,*]

[1]Bioinformatics Division, Institute of Cytology and Preventive Oncology, I-7 Sector-39, Noida-201301 and [2]Bioinformatics Centre, Institute of Microbial Technology, Sector 39A, Chandigarh 160036, India

Received August 15, 2012; Revised October 5, 2012; Accepted October 8, 2012

ABSTRACT

Plant-derived molecules have been highly valued by biomedical researchers and pharmaceutical companies for developing drugs, as they are thought to be optimized during evolution. Therefore, we have collected and compiled a central resource Naturally Occurring Plant-based Anti-cancer Compound-Activity-Target database (NPACT, http://crdd.osdd.net/raghava/npact/) that gathers the information related to experimentally validated plant-derived natural compounds exhibiting anti-cancerous activity (*in vitro* and *in vivo*), to complement the other databases. It currently contains 1574 compound entries, and each record provides information on their structure, manually curated published data on *in vitro* and *in vivo* experiments along with reference for users referral, inhibitory values (IC_{50}/ED_{50}/EC_{50}/GI_{50}), properties (physical, elemental and topological), cancer types, cell lines, protein targets, commercial suppliers and drug likeness of compounds. NPACT can easily be browsed or queried using various options, and an online similarity tool has also been made available. Further, to facilitate retrieval of existing data, each record is hyperlinked to similar databases like SuperNatural, Herbal Ingredients' Targets, Comparative Toxicogenomics Database, PubChem and NCI-60 GI_{50} data.

INTRODUCTION

Cancer is a well recognized global health problem responsible for ~7.6 million deaths (~13% of all deaths) worldwide, which is expected to rise to 13.1 million by 2030 (http://www.who.int/mediacentre/factsheets/fs297/en/index.html). Despite the progress in the field of cancer research, both developing and developed countries are in the grip of this deadly disease, and still there is a need to discover and develop anti-cancer therapeutic agents. It has long been recognized that natural products represent the richest source of high chemical diversity, providing the basis for identification of novel scaffold structures that serves as starting points for rational drug design (1). This can be one of the reasons that efforts have been directed to discover promising cancer therapeutic agents from natural sources. Over the years, many natural product-based drugs have been introduced in the market. According to a recent review, ~49% of drugs were either natural products or their derivatives that are used in cancer treatment (2). Moreover, between the year 2005 and 2010, 19 natural product-based drugs have been approved, among which 7, 10 and 2 have been classified as natural product (NP), semi-synthetic NPs and NP-derived drugs, respectively. Of these, five drugs, temsirolimus, trabectedin, ixabepilone, everolimus and romidepsin, have been developed in the area of oncology from 2007 to 2009 (3). Indeed, it has been suggested that less than one-fifth of the ring systems found in natural products are represented in current trade drugs (1).

As a result, during the past decades, several investigators have undertaken bioactivity-guided fractionation for various plant extracts and have isolated and/or identified a variety of bioactive compounds using *in vitro* cell-based cytotoxicity assays that exhibit anti-cancerous activity. Also, with advancement in discovery of target-based therapeutics, target/mechanism-based bioassays have been applied to identify potential mechanism of action of anti-tumour agents. Additionally, in few cases, *in vivo* evaluations have been undertaken to determine the efficacy of the compound of interest. This has resulted in generation of enormous data and revelation of hundreds of compounds that exhibit anti-cancerous activity against various cancers (4). Thus, providing researchers important resource to potentially identify the unique scaffolds and explore the molecular mechanisms. Despite presence of voluminous biomedical literature in PubMed that

*To whom correspondence should be addressed. Tel: +91 120 2579471/72; Fax: +91 120 2579473; Email: smagarwal@yahoo.com

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

provides evidence for variety of naturally occurring compounds demonstrating anti-neoplastic activity, to our knowledge, no resource exists that focuses on plant-based naturally occurring compounds with anti-cancerous activity, although a few natural compound databases like SuperNatural (5) and Herb Ingredients' Targets (6) or cancerous compound–target repository like CancerResource (7) have been published. SuperNatural is a large resource containing 3D structures and conformers of 45 917 purchasable natural compounds or its derivatives, whereas CancerResource is a cancer-specific repository that provides details of compound–target interactions derived from literature mining as well as from external sources, and Herb Ingredients' Targets database integrates information of 586 herbal ingredients (extracted from Chinese herbs) with protein-target information. Each of these is an excellent resource and has their own unique features, for example, in SuperNatural, a 3D superimposition facility is provided along with 3 million pre-computed conformers to account for flexibility during superimposition, whereas in CancerResource, two compounds can be matched by comparing growth activity profile across the NCI-60 cell lines. However, in these databases, information regarding the anti-cancerous *in vitro* and *in vivo* biological activity against various cancer cell lines has not been covered. Therefore, to complement these databases and to capture the intrinsic features of natural compounds that exhibit anti-tumour properties, we have designed and developed a central resource termed Naturally Occurring Plant-based Anti-cancer Compound-Activity-Target database (NPACT, http://crdd.osdd.net/raghava/npact/).

NPACT currently contains 1574 entries, and each record provides information on their structure, properties (physical, elemental and topological), experimentally determined *in vitro* and *in vivo* biological activity, description of cancer type, various cell lines, inhibitory values (IC_{50}, ED_{50}, EC_{50}, GI_{50}), molecular targets, commercial suppliers and drug likeness of compounds. NPACT concentrates on anti-cancerous natural compounds found in plants only. NPACT is unique in providing bioactivities of these natural compounds against different cancer cell lines and their molecular targets. It includes data extracted from PubMed and currently provides details for 353 cancer cell lines, which correspond to ~5214 compound–cell line interactions. It also provides information on protein targets that have been demonstrated in cancer cell lines to be inhibited by these naturally occurring compounds. At present, the database describes ~1980 experimentally validated compound–target interactions. Overall, NPACT is a specialized value added database that will enable exploration of biologically active cancer relevant plant-based naturally occurring compounds and will help in unravelling the scaffold diversity in the area of oncology.

DATA COLLECTION AND COMPILATION

To identify the plant-derived naturally occurring compounds with reported anti-cancerous activity, we searched PubMed and collected the relevant literature manually. A total of 181 journals pertaining to medicinal plant and natural product research were referred, resulting in a collection of 762 articles. The journals that have contributed significantly in the creation of the database include 'Journal of Natural Product', 'Phytochemistry', 'Cancer Research', 'Cancer Letters' and 'Biological and Pharmaceutical Bulletin', to name a few. After obtaining these articles, we read through the full text of each article to catalogue information like compound name, information pertaining to its *in vitro*/*in vivo* biological activity, that is, IC_{50}/ED_{50}/EC_{50}/GI_{50}, the cell line used for *in vitro* cytotoxicity assays, the model system in case of *in vivo* experiments and the protein target as documented in the references along with its tracking number (PMID). Overall, NPACT is divided into eight tables (i) general information table that contains all the basic information for a particular compound like its name, NPACT ID (unique), IUPAC, synonyms, PubChem ID, InChi, InChi key, SMART, CAS number and so forth; (ii) *in vitro* activity table that contains anti-cancer activity information along with inhibitory values against various cancer cell lines; (iii) *in vivo* activity table stores data for *in vivo* model system used, protein target and a brief remark explaining the observations as in the quoted reference; (iv) target table includes information of protein targets inhibited by the compound along with its inhibitory value; (v) cross reference table contains details regarding availability of our compound in other databases; (vi) property table contains information about physical, elemental and topological properties; (vii) filter table contains information about four computed filters (Lipinski's rule of five, mugge's filter, ghose filter and veber filter); and (viii) vendor table contains information about commercial availability of the compound (Figure 1).

DATABASE ARCHITECTURE AND WEB INTERFACE

Once all the information was gathered, we integrated the data in MySQL, an object-relational database management system (RDBMS), which works at the backend and the Web interface, and it was built in PHP, HTML and JavaScript as the front end. We have built NPACT on Apache HTTP server with MySQL server and PHP, HTML and JavaScript, as these are platform independent and are open-source software's/technology.

DATA ACCESS

The data in the NPACT can be easily accessed in a variety of ways. Users can query the database by using a simple text search tool that provides various options for searching like compound name, SMILE, InChi key, SMART, CAS number, class, PubChem ID and compound ID. The search results in the display of compound-centric information in a new page (Figure 2). The main page for each entry provides the following information: (i) general information, including the compound name, NPACT ID, IUPAC name, synonyms, class, InChi, InChi key, SMILES, SMART and CAS number; (ii) a schematic view of compound along with 3D representation that can be downloaded in MOL format; (iii) *in vitro* anti-cancer activity

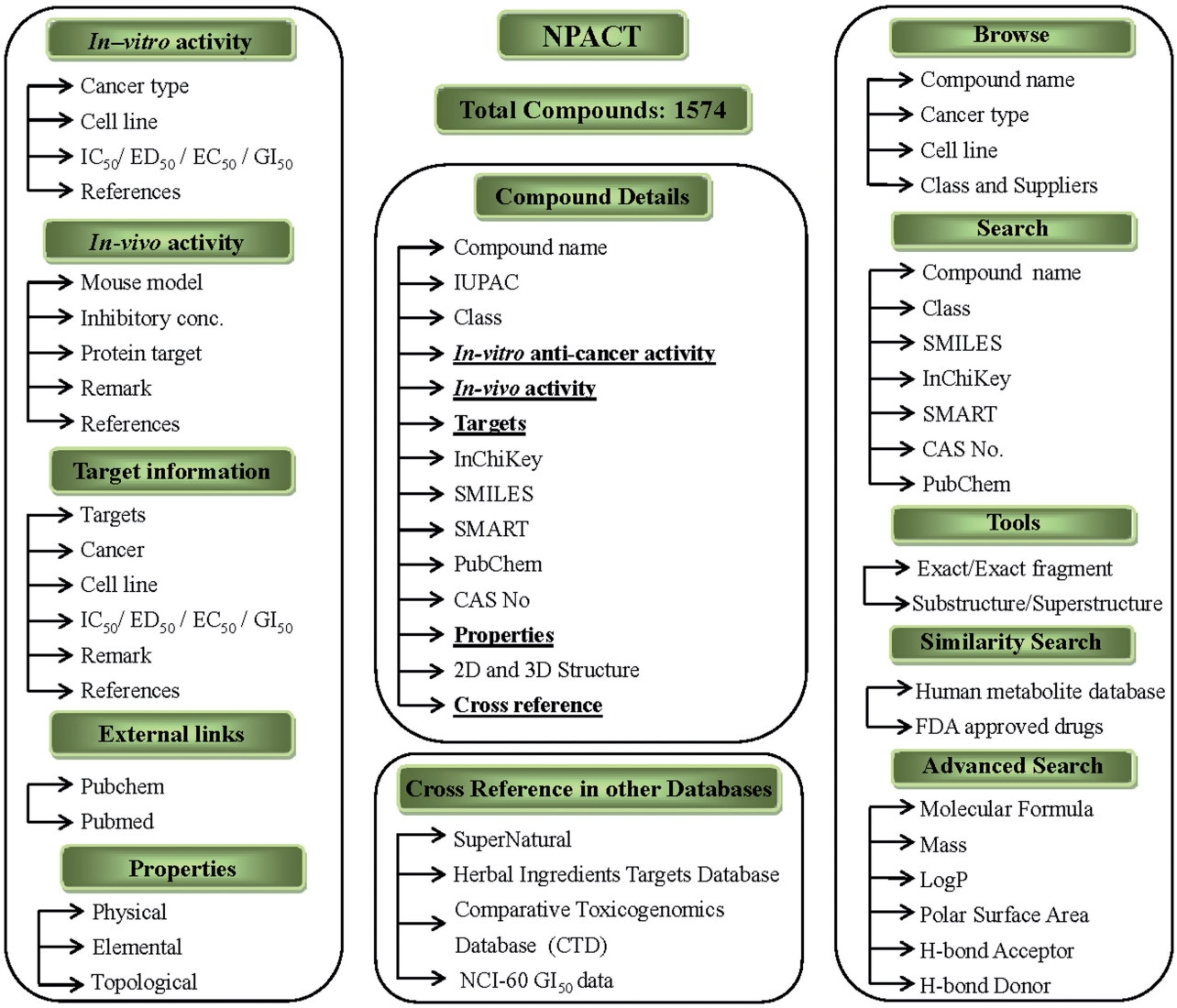

Figure 1. The architecture of NPACT database.

details; (iv) *in vivo* activity details; (v) it also provides information on protein targets that have been demonstrated in cancer cell lines to be inhibited by these naturally occurring compounds; (vi) hyperlink to various other similar databases like SuperNatural database, Herbal Ingredients' Targets database (HIT), Comparative Toxicogenomics database, NCI-60 GI_{50} data (vii) link to properties (physical, elemental and topological) (viii) drug likeness of compounds (ix) commercial availability/suppliers list; and (x) similarity search results against human metabolites as well as FDA approved drugs. Moreover, clicking on number of cell lines corresponding to *in vitro* anti-cancer link leads to another page that provides details of cell line, cancer type, the reported inhibitory value (IC_{50}/ED_{50}/EC_{50}/GI_{50}) and the corresponding reference from where the details were extracted. Similarly, by clicking YES corresponding to *in vivo* anti-cancer link, details of *in vivo* model, protein target, inhibitory concentration, remark and link to literature details for user's reference are disclosed. Further, a click on individual protein target reveals details about the experimental information of compound–target interactions.

Also, advance search option has been incorporated that provides the options to the user to select the compounds in a particular range on the basis of their physiochemical properties like molecular weight, XLogP, polar surface area, number of H-bond acceptor or donor and inhibitory concentration. NPACT offers a browsing section too, which allows the user to access the entire collection of compounds in five different ways, that is, by alphabetical order, cancer type, cell line, class name and supplier list. At present, NPACT covers 353 cell lines corresponding to 27 cancer types, 19 classes and 50 suppliers. The user can access the list of compounds for which the activity has been recorded for each of the cell lines as well as cancer type enabling him to identify compound of interest from the collection. The cancer-wise distribution of compounds in NPACT is shown in Figure 3. Thus, NPACT provides a gateway through which the cancer community working in

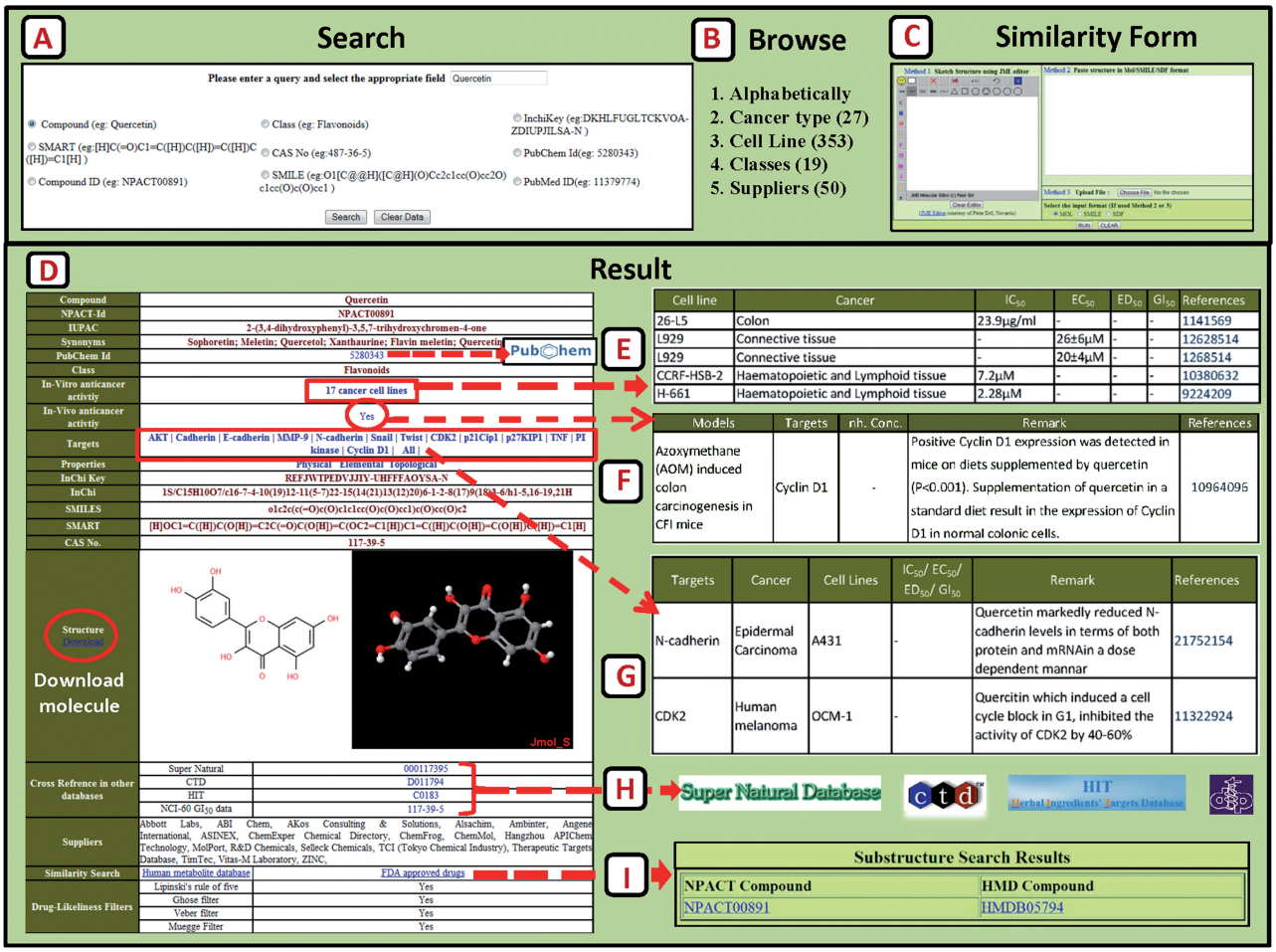

Figure 2. Schematic workflow of NPACT showing the three options Search (**A**), Browse (**B**) and Similarity search (**C**) by which a user can access the content in the database. Any of these methods leads to a compound-centric result page (**D**). Further linkage to *in vitro* activity information (**E**), *in vivo* (**F**), target (**G**), cross-links to other databases (**H**) and similarity result page (**I**) is shown.

the field of drug design and discovery can easily access NPACT for the latest information on the compounds exhibiting anti-cancerous activity.

SIMILARITY SEARCH

Apart from the data collected about anti-cancer activity, a simple Java based Java Molecular Editor tool (http://www.molinspiration.com/jme/) has been provided to draw structure to be searched in the database using JC search tool (8). Otherwise, the user can either paste or upload the structure as a MOL/SMILE/SDF file for undertaking similarity search. The tool provides exact substructure, exact fragment and super structure searching options.

Moreover, recent analysis suggests that drugs are on average more similar to endogenous metabolites (9,10); thus, it is expected that a compound similar to human metabolite will have higher chances to succeed as a drug. Therefore, we provide a link for each NPACT molecule that enables the user to determine the similarity relationship against FDA approved drugs as well as human metabolites.

ONLINE SUBMISSION TOOL

NPACT also offers an online submission facility to add plant-based natural compound entries that demonstrate anti-cancer properties but are not yet present in the database. The user can add new compound information by filling data submission form with specified fields (the fields indicated with star are mandatory) and entries will then be added to the database after validation.

DISCUSSION

During the past decade, a few natural compound databases like SuperNatural (5) and HIT (6) or cancerous compound–target repository like CancerResource (7) have been made available online. SuperNatural database is a resource containing 3D structures and conformers of 45 917 natural compounds, derivatives and analogues purchasable from different suppliers, whereas CancerResource is a database that integrates cancer-relevant relationships of compounds and targets, and HIT provides information for 586 herbal compounds and their corresponding protein targets. Thus, the

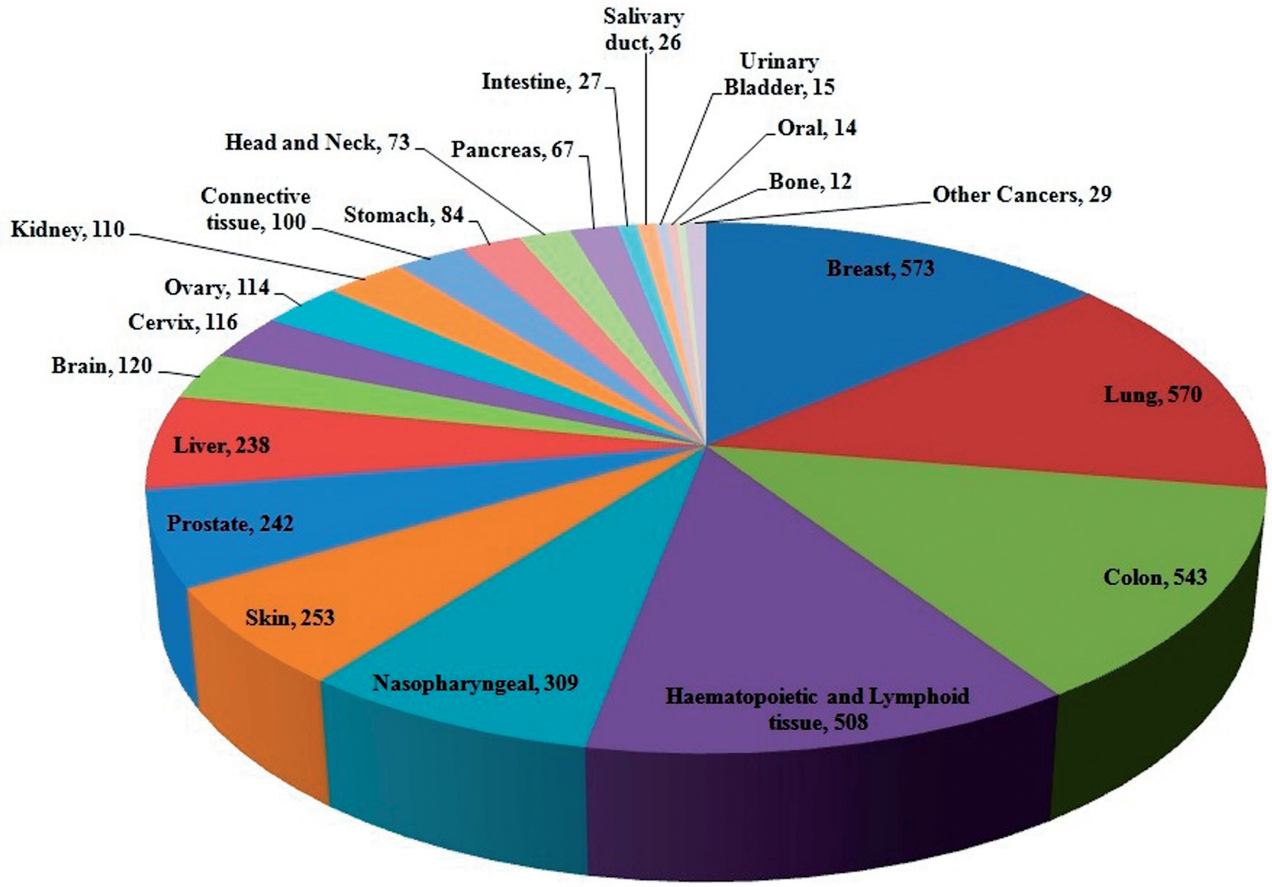

Figure 3. Cancer-wise distribution of the compounds in NPACT.

SuperNatural database is a general natural compound database of any origin without their corresponding biological activity, whereas the other two databases, that is, CancerResource and HIT have concentrated more on compiling protein-target information for the bioactive principle. In NPACT, we have compiled published experimental information on natural compounds found in plants only that exhibit anti-cancerous activity. NPACT is unique in providing *in vitro* bioactivities of these compounds against large number of cancer cell lines (IC_{50}/ED_{50}/EC_{50}/GI_{50}) as well as experimental information of *in vivo* studies demonstrating their anti-cancerous potential along with their molecular targets. Even in terms of number of compounds, the overlap between these existing repositories and NPACT is small. NPACT has 1574 compound entries, of which 1185 (75%), 1390 (88%) and 1417 (90%) are not found in CancerResource, SuperNatural and HIT data sets, respectively (Figure 4). Overall, there are as many as 1135 compounds (72%) that are not found in any of these databases. One of the reasons that the overlap between existing repositories and NPACT is small could be because of the reason that majority of entries in NPACT have been derived from literature published in recent years, that is, 84% of literature covered is post-2000 (Supplementary Figure S1). Thus, availability of NPACT as a public resource would furnish additional information with respect to the biological activity both *in vitro* and *in vivo* against various cancer cell lines. The authors think that it is worthy to have more than one comprehensive database in a field so as to provide alternate sources of information in case of unavailability or failures of any of the databases. Therefore, the present database will complement these existing databases in serving the scientific community.

This database enables users to identify compounds that either have activity against a particular or across large number of cancers/cell lines. For example, genistein (NPACT00605) has information for 28 cell lines corresponding to seven cancers. However, in all the four cell lines, BFTC-905, RT4, SCABER and TSGH8301 corresponding to urinary bladder cancer, the activity is $<5\,\mu g/ml$. This demonstrates that genistein is comparatively more effective against urinary bladder cancer as compared with other six cancers. Similarly, the molecule arnidiol (NPACT00278) has activity against 54 cell lines corresponding to nine cancers, out of which it is highly active ($GI_{50} < 10\,\mu M$) against 52 cell lines, demonstrating its across the board action against all the cancers tested.

Also, another advantage of this database will be that it would help in the process of drug discovery by providing researchers starting points for *in silico* screening of natural compounds as well as make available building blocks or scaffolds to be selected for the design of novel drugs. Moreover, comparative analysis of properties of synthetic,

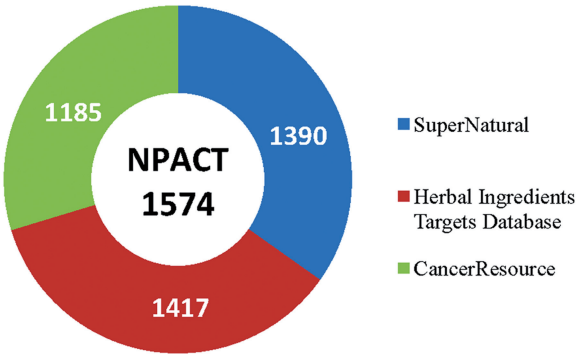

Figure 4. Number of compounds unique to NPACT in comparison with other databases.

natural compounds and drugs has revealed the various distinctness features of natural compounds (11). Thus, the data set offered by this database can also be used to identify structural features typical of natural products showing anti-cancerous activity. Further, the database can be of particular use for developing robust scaffold based quantitative structure–activity relationship models or various cancer cell line-based models as suggested in a recent study (12).

LIMITATIONS AND FUTURE DEVELOPMENTS

One major limitation in developing such databases is that extensive literature search is required to generate entries and expand the database. Also, we would like to mention that access to published literature acts as a hurdle, and projects of such nature cannot be undertaken without the community support. As an ongoing programme, we would like to include additional cell lines and *in vivo* data corresponding to the compounds already present in the database. This would enable to have a holistic picture for the compound against large number of cancers.

It is well known too that the absorption, distribution, metabolism, excretion and toxicity (ADMET) properties play an important role in the drug design process because failure of candidate compounds in the clinical phases is often associated with the properties like lack of efficacy, suboptimal formulation, toxicity or poor bioavailability. Therefore, efforts in future will be directed towards collecting experimental data published in literature related to ADMET for entries covered in NPACT. It is expected that this information will guide the selection of natural product scaffolds that are likely to have favourable pharmacokinetic properties and thus have greater likelihood for success in various phases of clinical trials.

Also, we propose to update this database regularly by adding new bioactive natural compounds from literature as well as incorporate data analysis tools that could potentially contribute in the development of novel therapeutic scaffolds/molecules.

SUMMARY

Researchers and pharmaceutical companies who are engaged in drug discovery process, in practice require *in vitro* and *in vivo* experimental data to take the decision regarding the *in vitro* potency and *in vivo* efficacy so as to make a considered decision regarding progressive optimization of leads to identify effective compounds with improved activity. Thus, the aim of developing NPACT database is to facilitate drug discovery in the area of cancer by providing a starting point.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figure 1.

ACKNOWLEDGEMENTS

The authors thank all their colleagues, friends and researchers who have provided them the published articles. They are also thankful to Open Source Drug Discovery (OSDD) for providing the platform to launch the website.

FUNDING

National Innovation Foundation grant [NIF/VARD/2010-11/43183]. Funding for open access charge: National Innovation Foundation, Department of Science and Technology (DST).

Conflict of interest statement. None declared.

REFERENCES

1. Koehn,F.E. and Carter,G.T. (2005) The evolving role of natural products in drug discovery. *Nat. Rev. Drug Discov.*, **4**, 206–220.
2. Newman,D.J. and Cragg,G.M. (2012) Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J. Nat. Prod.*, **75**, 311–335.
3. Mishra,B.B. and Tiwari,V.K. (2011) Natural products: an evolving role in future drug discovery. *Eur. J. Med. Chem.*, **46**, 4769–4807.
4. Chin,Y.W., Yoon,K.D. and Kim,J. (2009) Cytotoxic anticancer candidates from terrestrial plants. *Anticancer Agents Med. Chem.*, **9**, 913–942.
5. Dunkel,M., Fullbeck,M., Neumann,S. and Preissner,R. (2006) SuperNatural: a searchable database of available natural compounds. *Nucleic Acids Res.*, **34**, D678–D683.
6. Ye,H., Ye,L., Kang,H., Zhang,D., Tao,L., Tang,K., Liu,X., Zhu,R., Liu,Q., Chen,Y.Z. *et al.* (2011) HIT: linking herbal active ingredients to targets. *Nucleic Acids Res.*, **39**, D1055–D1059.
7. Ahmed,J., Meinel,T., Dunkel,M., Murgueitio,M.S., Adams,R., Blasse,C., Eckert,A., Preissner,S. and Preissner,R. (2011) CancerResource: a comprehensive database of cancer-relevant proteins and compound interactions supported by experimental knowledge. *Nucleic Acids Res.*, **39**, D960–D967.
8. Csizmadia,F. (2000) JChem: Java applets and modules supporting chemical database handling from web browsers. *J. Chem. Inf. Comput. Sci.*, **40**, 323–324.
9. Dobson,P.D., Patel,Y. and Kell,D.B. (2009) 'Metabolite-likeness' as a criterion in the design and selection of pharmaceutical drug libraries. *Drug Discov. Today*, **14**, 31–40.
10. Peironcely,J.E., Reijmers,T., Coulier,L., Bender,A. and Hankemeier,T. (2011) Understanding and classifying metabolite space and metabolite-likeness. *PLoS One*, **6**, e28966.
11. Ertl,P., Roggo,S. and Schuffenhauer,A. (2008) Natural product-likeness score and its application for prioritization of compound libraries. *J. Chem. Inf. Model.*, **48**, 68–74.
12. Bohari,M.H., Srivastava,H.K. and Sastry,G.N. (2011) Analogue-based approaches in anti-cancer compound modelling: the relevance of QSAR models. *Org. Med. Chem. Lett.*, **1**, 3.